Deep-coverage whole genome sequences and blood lipids among 16,324 individuals

Natarajan, Pradeep

2018-08-23

NHLBI TOPMED Lipids Working Grp 2018, 'Deep-coverage whole genome sequences and blood lipids among 16,324 individuals' Nature Communications, vol. 9, 3391. DOI: 10.1038/s41467-018-05747-8

http://hdl.handle.net/10138/243416
https://doi.org/10.1038/s41467-018-05747-8

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Deep-coverage whole genome sequences and blood lipids among 16,324 individuals


Large-scale deep-coverage whole-genome sequencing (WGS) is now feasible and offers potential advantages for locus discovery. We perform WGS in 16,324 participants from four ancestries at mean depth >29X and analyze genotypes with four quantitative traits—plasma total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol, and triglycerides. Common variant association yields known loci except for few variants previously poorly imputed. Rare coding variant association yields known Mendelian dyslipidemia genes but rare non-coding variant association detects no signals. A high 2M-SNP LDL-C polygenic score (top 5th percentile) confers similar effect size to a monogenic mutation (~30 mg/dl higher for each); however, among those with severe hypercholesterolemia, 23% have a high polygenic score and only 2% carry a monogenic mutation. At these sample sizes and for these phenotypes, the incremental value of WGS for discovery is limited but WGS permits simultaneous assessment of monogenic and polygenic models to severe hypercholesterolemia.

DOI: 10.1038/s41467-018-05747-8

1 Center for Genomic Medicine and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, USA. 2 Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. 3 Broad Institute of Harvard & MIT, Cambridge, MA 02142, USA. 4 Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. 5 Yale School of Medicine, New Haven, CT 06510, USA. 6 Department of Computational Biology & Bioinformatics, Yale University, New Haven, CT 06520, USA. 7 School of Medicine, University of Maryland, Baltimore, MD 21201, USA. 8 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA. 9 Department of Computational Medicine and Bioinformatics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA. 10 Society of Fellows, Harvard University, Cambridge, MA 02138, USA. 11 Department of Biological Chemistry, University of California, Los Angeles, CA 90095, USA. 12 Institute for Molecular Medicine Finland, Helsinki 00290, Finland. 13 Estonian Genome Center, University of Tartu, Tartu 51010, Estonia. 14 Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA. 15 Department of Biostatistics, University of Washington, Seattle, WA 98195, USA. 16 Department of Medicine, University of Mississippi Medical Center, Jackson, MS 35216, USA. 17 Sections of Preventive Medicine and Epidemiology and Cardiology, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. 18 Department of Epidemiology, Boston University School of Public Health, Boston, MA 02118, USA. 19 Framingham Heart Study, Framingham, MA 01702, USA. 20 Computer Science and Artificial Intelligence Lab (CSAIL), Massachusetts Institute of Technology, Cambridge, MA 02139, USA. 21 Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA. 22 Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 35216, USA. 23 Institute for Translational Genomics and Population Sciences, LABioMed and Departments of Pediatrics and Medicine, Harbor-UCLA Medical Center, Torrance, CA 90502, USA. 24 Departments of Human Genetics, Internal Medicine, and Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. These authors contributed equally: Pradeep Natarajan, Gina M. Peloso, Seyedeh Maryam Zekavat. These authors jointly supervised this work: Cristen J. Willer, Sekar Kathiresan. Correspondence and requests for materials should be addressed to S.K. (email: skathiresan1@mgh.harvard.edu). A full list of consortium members appears at the end of the paper.
Plasma lipids, including total cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides, are heritable risk factors for atherosclerotic cardiovascular disease. Understanding the inherited basis for plasma lipid levels has led to new treatments and to tests to identify individuals at risk for disease. Advances in technologies to characterize DNA sequence variants (i.e., Sanger sequencing, genotyping arrays, exome sequencing) have progressively allowed us to solve monogenic forms of dyslipidemia and to uncover common DNA sequence variants as well as rare mutations that contribute to plasma lipid levels in the population. However, due to the inherent limitations of genotyping arrays and exome sequencing, the non-coding regions of the genome remain incompletely characterized, particularly for rare mutations. In addition, the relative contribution of common DNA sequence variants and rare coding mutations to extreme lipid values in the population has not been delineated.

It is now possible to directly enumerate the whole-genome sequences of a large number of individuals. When performed at sufficient depth of coverage (>20-fold coverage per base), whole-genome sequencing (WGS) can detect single nucleotide polymorphisms (SNPs), insertions, and deletions across the allele frequency spectrum in both non-coding and coding regions. These advances allow us to test the incremental value of WGS as a tool for locus discovery and also develop a framework to understand why a specific individual might have an extreme lipid value. Toward these two goals, we studied the whole-genome sequences in 16,324 participants of European, African, East Asian, and Hispanic ancestries with available plasma lipids phenotypes.

In common variant association analyses, we replicate prior loci but detect newly associated variants not previously detected by prior genome-wide genotyping arrays or imputation. Analyses of rare coding variants yield known Mendelian dyslipidemia genes. Four approaches for analyzing rare non-coding variant associations do not detect any signals. WGS analysis of severe hypercholesterolemia shows a ten-fold enrichment of a high polygenic LDL-C score versus monogenic mutation for severe hypercholesterolemia. While the incremental value for WGS for locus discovery currently is limited largely due to relatively smaller sample sizes, WGS markedly improves the diagnostic yield of severe hypercholesterolemia through simultaneous assessment of monogenic and polygenic models.

Results
Deep-coverage WGS of 16,324 participants. Participants of the Framingham Heart Study (FHS), Old Order Amish (OOA), Jackson Heart Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), FINRISK Study (FIN), and Estonian Biobank (EST) underwent WGS (Fig. 1). Following quality control (Supplementary Table 1), 16,324 participants with plasma lipids available were included in the analysis (Supplementary Table 2). The mean (standard deviation (SD)) age was 51 (15) years and 8669 (53%) were women. About 5911 (36%) of the participants were of non-European ancestry (Supplementary Table 2, Supplementary Fig. 1a-c. The proportion of individuals on lipid-lowering medications was low (9%).

WGS target coverage was >30X for FHS, OOA, JHS, and MESA (as a part of the NIH/NHLBI Trans-Omics for Precision
We used a conventional statistical threshold for genome-wide associations among 16 loci reaching \( P \leq 5 \times 10^{-8} \) for LDL-C, HDL-C, and triglycerides in the FHS/OOA/JHS (TOPMed Phase I) variant call file (VCF). While only four (25%) loci displayed evidence of allelic heterogeneity at \( P < 5 \times 10^{-8} \), 13 (81.3%) had at least moderate evidence \( (P < 1 \times 10^{-6}) \) of allelic heterogeneity across the different ethnic groups available (Supplementary Table 8). Through conditional analyses for LDL-C, we identified a low-frequency haplotype specific to African Americans (MAF 0.1% FHS, 0% OOA, 1.0% JHS), including variants in \( LDLR \) (rs17249141), and enhancer 4 kb upstream from the \( LDLR \) promoter (rs17249141), and enhancer 4 kb upstream from the \( LDLR \) transcription start site (TSS) (rs114197570) (Supplementary Fig. 5, Supplementary Fig. 6). Presence of these variants resulted in a 28 mg/dl lowering of LDL-C \( (P = 2 \times 10^{-11}) \), suggesting increased expression of \( LDLR \) for carriers of the minor allele (Supplementary Fig. 7).

**Rare variant association study of coding variants.** To improve the power of detecting rare variant associations, we aggregated putative disruptive rare variants in coding sequences of each gene and tested the quantitative trait distribution among carriers of a set versus non-carriers. We aggregated coding sequence variants within each gene that were predicted to lead to loss of function (e.g., nonsense, canonical splice-site, or frameshift) or annotated as “disruptive” by the ensemble MetaSVM in silico
Rare variant association study of non-coding variants. Next, we sought to determine whether rare variants in non-coding regions associate with plasma lipids. We used four approaches to aggregate rare, non-coding variants. (Fig. 3). First, we aggregated variants within “sliding windows” of 3 kb in length\textsuperscript{13,14}. Second, we connected a non-coding variant to a gene if it resided in a segment annotated as an enhancer (and within 20 kb of a gene) or a region annotated as a promoter (and within 5 kb of the TSS of a gene). Third, using gene expression information, we connected a non-coding variant to a gene if it resided in a region annotated as an enhancer. Finally, we connected a non-coding variant to a gene based on a model which predicted gene-enhancer pairs using a chromatin-state model, including both HK27ac and Hi-C contact data, that we previously described\textsuperscript{15}. Regulatory annotations were derived from the ENCODE and NIH Roadmap projects for two cell types—HepG2 and adipose nuclei—relevant to lipoprotein metabolism. For these analyses, we considered a P < 0.05 / 254,032 groups = 2.0 × 10\textsuperscript{−7} as significant (Supplementary Table 10, Supplementary Table 11).

Using the sliding window approach to non-coding burden tests, we observed suggestive associations for 3 kb windows at the CETF (start chr16:56667000) locus (minimum P = 4 × 10\textsuperscript{−6}) and at the APOA1-C3-A4-A5 (start chr1:117094500) locus (minimum P = 8 × 10\textsuperscript{−6}) with HDL-C. A total of 17.6% of non-coding sliding windows occurring within 1 Mb of known lead lipid variants were at least nominally (P < 0.05) associated with lipids versus 4.4% in other regions of the genome across all traits (P difference = 8 × 10\textsuperscript{−22}). An aggregation of rare non-coding variants at only two genes—LDLR and APOE—were associated with LDL-C and total cholesterol (P < 2 × 10\textsuperscript{−7}) (Supplementary Fig. 9) (Supplementary Table 12). The strongest LDLR signal (P = 9.7 × 10\textsuperscript{−11}) was seen for an analysis that connected enhancers and promoters to a gene based on physical proximity (approach #2 above). Closer inspection of the specific variants shows that this signal is driven by the low-frequency haplotype specific to African Americans also detected with single variant association (Supplementary Fig. 10) (Supplementary Table 12). The strongest APOE signal (P = 8.1 × 10\textsuperscript{−26}) was observed in the model connecting enhancers to a gene by eQTLs for gene expression (approach #3 above). However, accounting for the strongest common variant association at the locus (rs7412, the APOE ε2 isofrom allele), this signal attenuates to non-significance (P = 1.8 × 10\textsuperscript{−2}), suggesting that the non-coding variants are driven by LD of the APOE ε2 isofrom. Beyond these two results, we found no additional signals for a burden of non-coding variants.

Contribution of mono- and polygenic models to extreme LDL-C. With the availability of sequence in both coding and non-coding regions in the same samples, we estimated the simultaneous contribution of monogenic and polygenic determinants to extreme LDL-C in a population-based sample of European (EA) and African (AA) ancestry. We defined “extreme” as the top or bottom 5th ancestry-specific percentile of LDL-C. Analyses were conducted in FHS and MESA-EA subjects (extreme cutoff as LDL-C >183 mg/dl or LDL-C <72.9 mg/dl) and JHS and MESA-AA subjects (extreme cutoff as LDL-C >198.6 mg/dl or LDL-C <71 mg/dl), separately.

Among participants with extremely high LDL-C, we searched for mutations in any of six Mendelian genes previously implicated as causing elevated LDL-C (LDLR, APOB, PCSK9, ABCG5, ABCG8, and LDLRAP1) (Supplementary Table 13). To determine polygenic contribution, we implemented a systematic approach to derive, test, and validate a new “genome-wide” polygenic score for LDL-C using mutually independent datasets. A polygenic score provides a quantitative assessment of the cumulative risk associated with multiple common risk alleles for each individual.

We derived polygenic scores by three approaches: (1) only inclusion of genome-wide significant variants (P < 5 × 10\textsuperscript{−8} in separate discovery\textsuperscript{7}), (2) r\textsuperscript{2} and P value thresholds to restrict variants without rescaling weights, and (3) entire summary results of 2M variants (LDPred) with rescaled weights based on r\textsuperscript{2} and P values\textsuperscript{16}. We derived polygenic scores based on the association statistics of all available common (MAF > 0.01) SNPs with LDL-C, as determined by our previously published genome-wide association study\textsuperscript{7}.

As a baseline, we generated an additional polygenic score restricted to lead variants (P < 5 × 10\textsuperscript{−8}) at distinct genomic loci, weighted by discovery estimated effects (“restricted score”). Second, we applied various r\textsuperscript{2} and P value thresholds to the previously published results. Finally, we used the LDPred computational algorithm which constructs genome-wide polygenic scores across full summary statistics\textsuperscript{16}. Prior simulations have suggested that approaches additionally including variants with sub-genome-wide significance may improve the predictive capability of polygenic risk scores\textsuperscript{17}. To include such variants, LDPred re-weights corresponding per-variant weights from our prior genome-wide association study\textsuperscript{7} based on LD, discovery P values, and a range of estimated causal fraction (e.g., non-zero effect sizes) markers. The correlation between the variants was assessed using the European reference population from the 1000 Genomes study\textsuperscript{17}. The best score was determined based on maximal model fit (R\textsuperscript{2}) from a linear regression models in a health-care biobank of 25,534 unrelated individuals (Nord-Trøndelag Health Study, HUNT)\textsuperscript{18} (Supplementary Table 14).

For LDL-C, a genome-wide polygenic score incorporating 2 million SNPs with LDPred provided the best model fit (Supplementary Table 15). Compared to a restricted score of 59 SNPs independently significant associated with LDL-C, a relative increase of 21.6% of LDL-C variance was explained by the expanded 2M-SNP score (r\textsuperscript{2} restricted = 0.245 vs. r\textsuperscript{2} expanded = 0.298). We applied this polygenic score separately within the WGS samples in FHS, JHS, and MESA. We labeled individuals as having a high polygenic score if they fell in the top 5th percentile of race-specific score distributions (Tables 1 and 2).

Among EA participants, a monogenic mutation was associated with an odds ratio of 10.92 (95% CI 3.71(32.14) for extremely high LDL-C, whereas a high polygenic score associated with an odds ratio of 7.65 (95% CI 5.56–10.52). In EA individuals, those who carried a monogenic mutation had 30 mg/dl higher LDL-C (when compared with non-carriers; P = 2.1 × 10\textsuperscript{−4}) and those who had a high polygenic score had 33 mg/dl greater LDL-C (when compared with all others; P = 1.7 × 10\textsuperscript{−7}). Of the 287 EA participants with extremely high LDL-C, 2% carried a monogenic mutation and 23% had a high polygenic score.
Fig. 3 Schematic of non-coding rare variant analyses. Four grouping schematics of rare non-coding variants (MAF <1%). (1) The sliding window approach tiles across the genome at fixed widths, only including variants overlying annotations consistent with enhancers, promoters, and DHS in non-exonic regions. All other approaches attempt to map non-coding putative functional genomic regions with discrete genes as the analytical unit. Overall, they are based on: (2) promoter, enhancer, and DHS annotations near a gene’s transcription start site, (3) co-occurrence of enhancer and DHS annotations with HepG2 gene expression, and (4) H3K27ac marks within Hi-C contact regions mapped to genes. DHS DNase hypersensitivity site, MAF minor allele frequency.
Among AA participants, a monogenic mutation was associated with an odds ratio of 7.43 (95% CI 3.01–18.35) for extremely high LDL-C, whereas a high polygenic score was associated with an odds ratio of 3.2 (95% CI 2.46, 5.58) for extremely low LDL-C. In AA individuals, those who carried a monogenic mutation had 41 mg/dl higher LDL-C (when compared with all others; $P < 2 	imes 10^{-16}$) for extremely high LDL-C and 23.2 mg/dl higher LDL-C (95% CI 15.0–31.5; $P = 3.8 	imes 10^{-8}$) for extremely low LDL-C compared with individuals without a high polygenic score. Among ARIC-AA (N = 1907) participants, a high polygenic score was associated with an odds ratio of 6.25 (95% CI 6.2, 76.15) for extremely high LDL-C and 3.01 (95% CI 2.1, 4.89) for extremely low LDL-C compared with individuals without a high polygenic score.

We analyzed the monogenic and polygenic contribution to extremely low LDL-C in EA and AA participants and found similar patterns where monogenic mutations as well as a polygenic score conferred similar effect sizes (Tables 1 and 2).

**Discussion**

We performed WGS in 16,342 ethnically diverse individuals and analyzed the incremental value of WGS for locus discovery.
for blood lipid levels and for clinical interpretation. We replicated associations for 28 common variant loci previously associated with lipids in much larger genome-wide association analyses. We identified an association for a low frequency 1-bp deletion at 9p24.1 with HDL-C. We replicated burden associations of rare coding mutations at known Mendelian lipid genes. However, we did not detect any burden associations of rare non-coding mutations through four different approaches.

Lastly, we developed a genome-wide polygenic score and showed that such a score confers an effect size on LDL-C similar to carrying a monogenic mutation and is present in tenfold more individuals with severe hypercholesterolemia than monogenic mutations. At these sample sizes and for these phenotypes, the incremental value of WGS as a discovery tool was limited but WGS allowed us to simultaneously assess the contribution of monogenic and polygenic models to severe hypercholesterolemia.

These results permit several conclusions. Using WGS as a discovery tool, the incremental yield of new loci was modest. Current sample sizes for WGS are much smaller compared to genome-wide association and whole exome sequencing studies clearly limiting relative power for detecting associations for common/low-frequency non-coding variants and rare coding variants. Despite genome-wide interrogation of rare variant signals in non-coding space, we identified no burden-of-rare-variant signals using four different aggregation approaches and regulatory annotations from two relevant tissues.

Mutation target size and natural selection pressure are smaller in non-coding regions when compared with coding regions based on these considerations, power calculations have suggested that sample sizes may need to be considerably larger to identify rare variant burden associations in non-coding regions compared to coding regions. While sample size is an important determinant of power, prioritization of putative causal rare non-coding variants remains a major power limitation. Functional annotations from reference datasets largely prioritize functional sequence and MAF thresholds assist in prioritizing causal variants, but this likely retains a large fraction of benign variants. Genome-wide organism-level functional variant scores offer the promise of improved prioritization but did not improve associations at LDLR and APOE. Novel, genome-wide tissue-level functional scores may improve prioritization compared to organism-level scores. Assessments of consequence for rare coding mutations in experimental systems has improved associations of lipid-related genes beyond in silico tools. Similar systematic approaches for rare non-coding variants in relevant tissues may further improve power.

WGS in diverse populations permits discovery of novel associated variants. Most of the observed lead single variant associations at known loci were previously tagged by lead variants from genome-wide association analyses of largely European ancestry participants. Our trans-ethnic analyses yielded new lead variants at one-third of known lipid loci not previously tagged by prior lead loci. Additionally, variant classes not previously detected by array-genotyping and whole-exome sequencing are associated with lipids. We observed that a 1-bp deletion, not correlated with previously cataloged variants, was associated with HDL cholesterol. These observations indicate that new variants are detected not only by including diverse ethnicities, but also WGS can overcome many limitations of imputation for variant discovery, including application in non-Europeans, variable coverage in genome-wide genotype arrays, and detection of rarer variants.

Of great interest, we observed that the relative contribution of polygenic score to extremely high LDL-C is considerably greater than monogenic mutations. For example, in EA individuals, both high polygenic score and a monogenic mutation confer similar effects (~30 mg/dl higher LDL-C) but a high polygenic score is present in 20% of participants with extremely high LDL-C whereas a monogenic mutation is present in only 2%. In most individuals who carry diagnosis of familial hypercholesterolemia, no monogenic mutation is identified with clinical exome sequencing for a large fraction of these “mutation-negative” familial hypercholesterolemia, high polygenic scores may be operative. WGS permits the application of simultaneous assessment of monogenic determinants as well as the most optimally performing polygenic score with relative ease.

Our observed monogenic carrier rates for severe hypercholesterolemia (2%) are consistent with observations in other population-based cohorts and health-care-associated biobanks but lower than for patients with clinical criteria for familial hypercholesterolemia (up to 24%) particularly those clinically referred for familial hypercholesterolemia genetic testing (up to 50%). As anticipated, this subgroup is also likely to have a greater monogenic relative to polygenic contribution.

Important limitations should be considered. First, appropriate definitions of statistical significance for WGS association analyses have not been harmonized in the field. The convention of $\alpha = 5 \times 10^{-8}$ comes from the assumption of performing 1,000,000 independent tests. Based on our findings and simulations from others, may be more appropriate for analyses across diverse ethnicities to allele frequency 0.1%. Second, power is somewhat diminished with our rare variant meta-analysis approach to combine $P$ values with Fisher’s method. Given known diverse coding mutations within Mendelian genes with bidirectional effects and the inability to assume unidirectional effects within the non-coding space, we employed a SKAT statistical framework. Prior approaches leveraging covariance matrices for SKAT meta-analysis were computationally inefficient for the dataset and multiple grouping strategies. Thus, our approach is conservative. Third, the polygenic scores described here were derived from genome-wide association studies performed largely in EA ancestry participants. Because allele frequencies, LD patterns, and effect sizes of common polymorphisms vary by ancestry, the predictive capacity of polygenic score was attenuated in non-European ancestry individuals. Furthermore gene flow between ancestral groups and resultant admixture for an individual further hinders accuracy of polygenic risk scores derived from distinct populations for application at the individual level. This is an important limitation for the field that requires efforts to characterize common genomic variation influencing complex traits among non-Europeans and develop locus admixture-aware polygenic risk scoring.

In summary, we present a large-scale WGS analysis of plasma lipids in 16,324 ethnically diverse participants. Common, non-coding variants and rare, coding variants contribute to plasma lipid variation; however, association signals for rare, non-coding mutations were not detectable. Among participants with severe hypercholesterolemia, a high polygenic score was present in tenfold more individuals than a monogenic mutation.

**Methods**

**Study participants.** Study participants were from the FHS (N = 4064), JHS (N = 3247), OOA (N = 1083), MESA (N = 4510), FIN (N = 1165), and the EST (N = 2381). Each study was previously approved by respective institutional review boards (IRBs), including for the generation of WGS data and association with phenotypes. All participants provided written consent. The analyses of WGS data with plasma lipids was approved by the Massachusetts General Hospital IRB (MGH IRB 2016P001308). Please refer to Supplementary Note 1 for study participant details.

**WGS, variant calling, and genotyping.** Sequencing was performed at one of the four sequencing centers, with all members within a cohort sequenced at the same center. For the TOPMED phase 1 data, 4148 FHS individuals and 1905 OOA individuals were sequenced at the Broad Institute of Harvard and MIT (Cambridge, Massachusetts, USA).

**Statistical methods.** The association analysis of WGS data focused on LDL-C. WGS was performed at the Broad Institute of Harvard and MIT (Cambridge, Massachusetts, USA). The dataset comprised 16,324 individuals with blood lipids and WGS data. The analysis focused on LDL-C associations, with additional analyses performed for HDL-C and triglycerides. The study aimed to identify new rare and common variants associated with LDL-C, using a polygenic score approach and a monogenic mutation approach.

**Polygenic score.** The polygenic score was calculated using a linear regression model, with LDL-C as the outcome variable and genotype as the predictor. The model included all SNPs with minor allele frequency (MAF) greater than 0.01 and with a call rate greater than 0.9. The polygenic score was calculated as the sum of the estimated effect sizes for each SNP, weighted by the SNP allele frequency. The polygenic score was normalized to have a mean of 0 and a standard deviation of 1, allowing for comparison across different studies.

**Monogenic mutation.** Monogenic mutations were identified using a targeted approach, focusing on Mendelian lipid genes. The approach included a literature review, expert review, and computational analysis. The targeted mutations were confirmed using Sanger sequencing and functional validation studies. The effect of each mutation was evaluated using in vitro and in vivo functional assays.

**Conclusion.** The study presented a large-scale WGS analysis of plasma lipids in 16,324 ethnically diverse participants. Common, non-coding variants and rare, coding variants contribute to plasma lipid variation; however, association signals for rare, non-coding mutations were not detectable. Among participants with severe hypercholesterolemia, a high polygenic score was present in tenfold more individuals than a monogenic mutation. The study highlighted the importance of rare variants in Mendelian genes and the potential role of polygenic risk scores in lipid variation.

Methods

**Study participants.** Study participants were from the FHS (N = 4064), JHS (N = 3247), OOA (N = 1083), MESA (N = 4510), FIN (N = 1165), and the EST (N = 2381). Each study was previously approved by respective institutional review boards (IRBs), including for the generation of WGS data and association with phenotypes. All participants provided written consent. The analyses of WGS data with plasma lipids was approved by the Massachusetts General Hospital IRB (MGH IRB 2016P001308). Please refer to Supplementary Note 1 for study participant details.

**WGS, variant calling, and genotyping.** Sequencing was performed at one of the four sequencing centers, with all members within a cohort sequenced at the same center. For the TOPMED phase 1 data, 4148 FHS individuals and 1905 OOA individuals were sequenced at the Broad Institute of Harvard and MIT (Cambridge, Massachusetts, USA).

**Statistical methods.** The association analysis of WGS data focused on LDL-C. WGS was performed at the Broad Institute of Harvard and MIT (Cambridge, Massachusetts, USA). The dataset comprised 16,324 individuals with blood lipids and WGS data. The analysis focused on LDL-C associations, with additional analyses performed for HDL-C and triglycerides. The study aimed to identify new rare and common variants associated with LDL-C, using a polygenic score approach and a monogenic mutation approach.

**Polygenic score.** The polygenic score was calculated using a linear regression model, with LDL-C as the outcome variable and genotype as the predictor. The model included all SNPs with minor allele frequency (MAF) greater than 0.01 and with a call rate greater than 0.9. The polygenic score was calculated as the sum of the estimated effect sizes for each SNP, weighted by the SNP allele frequency. The polygenic score was normalized to have a mean of 0 and a standard deviation of 1, allowing for comparison across different studies.

**Monogenic mutation.** Monogenic mutations were identified using a targeted approach, focusing on Mendelian lipid genes. The approach included a literature review, expert review, and computational analysis. The targeted mutations were confirmed using Sanger sequencing and functional validation studies. The effect of each mutation was evaluated using in vitro and in vivo functional assays.

**Conclusion.** The study presented a large-scale WGS analysis of plasma lipids in 16,324 ethnically diverse participants. Common, non-coding variants and rare, coding variants contribute to plasma lipid variation; however, association signals for rare, non-coding mutations were not detectable. Among participants with severe hypercholesterolemia, a high polygenic score was present in tenfold more individuals than a monogenic mutation.
MA), while 3266 JHS individuals were sequenced at University of Washington Northwest Genomics Center (Seattle, WA). About 4601 MESA individuals were additionally sequenced at the Broad Institute of Harvard and MIT (Cambridge, MA). Three separate callsets were utilized due to timeline of availability as well as data restrictions.

TOPMed phase 1 BAM files provided by the sequencing centers were harmonized by the TOPMed Informatics Research Center (IRC) before joint variant discovery and genotype calling across studies. In brief, sequence data were received from each sequencing center in the form of bam files mapped to the 1000 Genomes hg19 reference sequence. Sequence processing was coordinated among the sequencing centers by the “GotCloud” processing pipeline. The two sequence quality criteria used in order to pass sequence data on for joint variant discovery and genotyping are: estimated DNA sample contamination below 3%, and fraction of the genome covered at least 10× 95% or above. DNA sample contamination was estimated from the sequencing center read mapping using software varantiago. The genotype callsets used for analysis are from “freeze 3a” of the variant calling pipeline performed by the TOPMed IRC (Center for Statistical Genetics, University of Michigan, Hyun Min Kang, Tom Blackwell, and Goncalo Abecasis). The software tools used in the generation of the pipeline are available in the following repository: https://github.com/stagten/topmed_freeze3_calling. Variant detection (SNPs and indels) from each sequenced (and aligned) genome is performed by vt discover2 software tool. The variant calling software tools are under active development; updated versions can be accessed at http://github.com/atks/vt or http://github.com/hyunminkang/apigenome.

WGS for MESA, FINRISK, and the Estonian Biobank was performed using the Illumina HiSeqX platform at the Broad Institute of Harvard and MIT (Cambridge, MA). DNA samples are informatically received into the Genomics Platform’s Laboratory Information Management System via a scan of the barcodes using a Biosero flatbed scanner. All samples are then weighed on a BioMicro Lab’s XL20 to determine the volume of DNA present in the sample tubes. Following thin, the samples are quantified in a process that uses PICO-green fluorescent dye. Once volumes and concentrations are determined, the samples are then handed off to the Sample Retrieval and Storage Team for storage in a locked and monitored −20°C walk-in freezer.

Libraries were constructed and sequenced on the Illumina HiSeqX with the use of 151-bp paired-end reads for WGS and output was processed by Picard to generate aligned BAM files (to hg19). Samples were tracked by automated LIMS messaging. Samples were fragmented with acoustic shearing and libraries were prepared with a KAPA Biosystems kit. Libraries were normalized to 1.7 nM. Cluster amplification was performed using Illumina cBot and flowcells were sequenced in HiSeq X. Variants (SNPs and indels) were discovered using the Geome Analysis Toolkit (GATK) v3 HaploTypeCaller according to Best Practices. VCFs from MESA samples were generated in one callset. Finland and Estonia samples were jointly called in a separate callset.

Whole-genome sequence quality control. The following three approaches were used by the TOPMed Genetic Analysis Center to identify and resolve sample identity issues: (1) concordance between annotated sex and biological sex inferred from the WGS data, (2) concordance between prior SNP array genotypes and WGS-sequenced genotypes, and (3) comparisons of observed and expected relatedness from pedigrees. The variant filtering in TOPMed Freeze 3 were performed by (1) first calculating Mendelian consistency scores using known familial relatedness and duplicates and (2) training SVM classifier between the known variant sites (positive labels) and the Mendelian inconsistent variants (negative labels). Two additional hard filters were applied: (1) Excess heterozygosity filter (EXHET), which filtered out by this filter, and (2) Mendelian discordance filter (DISC), or 3 or more Mendelian inconsistencies on a given variant was filtered out if this filter. Functional annotation for each variant was provided in the INFO field using snpEff 4.1 with a GRCh37.75 database. Analysis used hard-call genotypes, without genotype likelihoods. Genotypes with a depth <10 were excluded.

Additional measures for quality control of TOPMed Phase 1 Freeze 3 and quality control for MESA, Finland, and Estonia were performed using the Hail software package (https://hail.is). Samples were filtered by contamination (>3.0% for all, except >5.0% for Finland and Estonia), chimeras >5%, GC dropout >4, raw coverage (<30X for all, except <19X for Finland and Estonia), indeterminant genotypes (sex or SNP not called), and low phasing confidence (10% or less for each cohort) were filtered out. Variants for MESA, Finland, and Estonia were initially filtered by GATK Variant Quality Score Recalibration. Additionally, genotypes with GQ <20, DP < 10 or >200, and poor allele balance (homozygous with <0.90 supportive reads or heterozygous with <0.20 supportive reads) were removed. And variants within low quality region were removed across all three sequencing centers with 5% missing calls, quality by depth <2 (SNPs) or <3 (indels), InbreedingCoeff < −0.3, and pHEW < 1 x 10−8 (within each cohort) were filtered out.

Annotation. Variants were annotated with Hall using annotations from Ensemble’s Variant Effect Predictor for protein-coding annotations and RegMap20, HoneyBadger2-intersect release. Conventionally measured plasma lipids, including total cholesterol, LDL-C, HDL-C, and triglycerides, were included for analysis. LDL-C was either calculated by the Friedewald equation when triglycerides were <400 mg/dl or directly measured. Given the average effect of statins, when statins were present, total cholesterol was adjusted by dividing by 0.7, as previously done. Triglycerides were natural log transformed for analysis. Phenotypes were harmonized by each cohort and deposited into the dbGaP TOPMed Exchange Area.

Common plus low-frequency variant association analysis. Single variant analysis was performed in EPACTS (https://genome.sph.umich.edu/wiki/EPACTS) with Efficient Mixed-Model Association eXpedited (EMMAX) for associating each variant with each lipid trait using a conditioning linear model. We made predictions separately for each of the 127 reference epigenomes (http://github.com/hyunminkang/apigenome). The genotype callsets used for analysis are from “freeze 3a” of the variant calling pipeline performed by the TOPMed IRC (Center for Statistical Genetics, University of Michigan, Hyun Min Kang, Tom Blackwell, and Goncalo Abecasis). Empirical kinship matrices were first generated for each VCF (“make-kin”) using default parameters. Next, association analyses (”single”) were performed adjusting for age, age2, sex, cohort, self-reported ethnicity (for MESA), and an empirically derived kinship matrix to account for both familial and more distant relations. Within each VCF, TOPMed performed the MAF > 0.01 variants with MAF > 0.1% were adjusted for by conditioning on MAF > 0.01. Variants were meta-analyzed across all three VCFs using METAL (https://genome.sph.umich.edu/wiki/METAL)47. Summary-level statistics only for variants with MAF > 0.1% for the given VCF were included in the meta-analysis. Statistical significance α of 5 × 10−8 was used for these analyses.

For loci with at least one variant with P < 5 × 10−8 within the TOPMed Phase 1 VCF, iterative conditional association analysis was performed. Iterative conditioning was performed until P < 1 × 10−8 was attained.

Rare variant association analyses. We first identified rare (MAF <1%) mutations for each VCF within the coding sequences. After Variant Effect Predictor annotation, we identified loss-of-function (e.g., nonsense, canonical splice-site, and frameshift) and disruptive missense (by MetaSV49 in canonical transcripts as described by Ensembl).

We further performed rare variant association tests within the non-coding regions (Supplementary Figure 7). As before, we performed a “sliding window” approach aggregating 3 kb (overlapping by 1.5 kb) windows and considering rare variants occurring within enhancer or promoter elements at DNase I hypersensitivity sites. For non-coding tests, we next attempted to link rare non-coding variants with genotype association testing using regulatory annotations for HepG2 and adipose nuclei from ENCODE and NIH Roadmap. Given prior observations showing enrichment of functional promoter variants at LIPG with HDL-C extremes, we similarly aggregated variants near TSSs. Prior studies have shown that approximately 80% of cis-eQTLs fall within 100 kb of TSS52. To increase the likelihood of mapping regulatory variants to the nearest gene, we were more restrictive and included variants overlapping promoter sequences ±5 kb and enhancer sequences ±20 kb of TSS at DNase I hypersensitivity sites.

We also linked chromatin state defined enhancers with genes using data from the Roadmap Epigenomics project53 and the method presented previously54 with a few small modifications. The method predicts links using chromatin the tracks information, position of the enhancer relative to the TSS, and the correlation of multiple chromatin marks with gene expression across cell types. Here we used gene expression with the correlation of the signal of five chromatin marks: H3K27ac, H3K9ac, H3K36me1, H3K4me2, and DNaseI hypersensitivity. The gene expression data were trimmed to remove outliers. The protein-coding rare variants from the Roadmap Epigenomics project (available in the file 57epigenomes.RPKM.pc from http://compbio.mit.edu/roadmap; Universal Human Reference was excluded). The chromatin mark signal was the log10(P) tracks averaged to a 200-bp resolution. As input to our code, we used the version of those tracks first averaged at 25-bp resolution using the ”Coarse with command of ChromInput. In computing correlation between a specific chromatin mark signal and gene expression, we used the Pearson correlation and omitted from the calculation samples lacking both chromatin mark signal and gene expression data. We made predictions separately for each of the 127 reference epigenomes and locations assigned to chromatin states, 6, ENCODE, 7, Enh, and 12, EnhBiov of the 15-state core 5-marks ChromHMM model52,54. We restricted our predictions to chromatin state assignments on chr1-22 and chrX. We considered linking 200-bp bins within 1 Mbp of a TSS of each gene as annotated in the file Ensembl_v65. Genecode_v10.ENSG_gene_info available from http://compbio.mit.edu/roadmap (ref. 53). If a gene had multiple TSS, then we used the gene with the most TSS.

The method for linking is based on determining for each combination of cell type, chromatin state, and position relative to the TSS the estimated probability the
set of correlations we observed would come from the actual data compared to randomized data. To this end, we created a training set of actual observed correlations (positive examples) and correlations computed after randomizing which gene expression values were assigned to which genes (negative examples) separately for each combination of cell type, chromatin state, and position relative to the TSS. Each entry in the training set has five features corresponding to correlations for each considered chromatin mark. There is a corresponding positive example and a corresponding negative example for each instance of the specified chromatin state in the specified cell type at the specified position relative to the TSS or within 5 kb of it (for smoothing purposes). We trained a logistic regression classifier to discriminate actual correlations with randomized correlations. We used the logistic regression library implemented in the Weka package version 3.7.3 with the regularization parameter set to 1.59. For considering linking a specific instance of a chromatin state assignment in a specific cell type and position relative to the TSS of a gene, we applied the corresponding classifier. Let $p$ denote the probability the classifier gives of being in the positive class of the actual observed correlations. We retained those links for which $p$ exceeded 0.25. The method used here is implemented in the code LinkingRM.java. For the analyses presented here, we used those links for the primary enhancer state, $7_{\text{Einh}}$.

To connect non-coding variants with putative target genes, we predicted functional gene-enhancer pairs using a chromatin state-based model we previously developed. This model assumes that the impact of an enhancer on gene expression is determined by the product of its intrinsic “Activity” (for which we use quantitative DNase-Seq and H3K27ac levels as a proxy) and the “Contact Frequency” at which the enhancer physically encounters its target promoter in the nucleus (for which we use Hi-C data as a proxy). We previously showed that Activity by Contact (ABC) model accurately identifies enhancers whose perturbation leads to changes in gene expression in the human MYC locus, and we have since found that the same model can identify enhancers across other gene loci and cell types (Fulco, C., Lander, E., and Ingreyt, J., in preparation). We extended our previously published model to predict polymorphisms in gene-specific connections in the liver, using DNase-Seq and H3K27ac ChIP-seq data from a hepatocarcinoma cell line (HePG2) previously generated by the ENCODE project. To define putative regulatory elements, we expanded DNase-Seq peak calls from ENCODE by 500 bp on either side and merged overlapping peaks. For each element, we calculated Activity as a function of the number of peak covers at H3K27ac and DNase-Seq. Because a high-resolution Hi-C data is not available for HePG2 cells, we estimated the Contact probability between putative regulatory elements and genes using the average profile across deeply sequenced Hi-C libraries from seven different cell types as previously described. For each putative enhancer-gene pair, we calculated an “ABC score” equal to the product of Activity × Contact across all other putative elements within 5 Mb of the target gene. We tuned free parameters in this model (such as the relative weight of DNase-Seq and H3K27ac data and a pseudocount to add to Hi-C data) and selected the parameters that optimize the sum of Activity × Contact across all other putative elements within 5 Mb of the target gene. We tuned free parameters in this model (such as the relative weight of DNase-Seq and H3K27ac data and a pseudocount to add to Hi-C data) and selected the parameters that optimize the sum of Activity × Contact across all other putative elements within 5 Mb of the target gene. We tuned free parameters in this model (such as the relative weight of DNase-Seq and H3K27ac data and a pseudocount to add to Hi-C data) and selected the parameters that optimize the sum of Activity × Contact across all other putative elements within 5 Mb of the target gene. We tuned free parameters in this model (such as the relative weight of DNase-Seq and H3K27ac data and a pseudocount to add to Hi-C data) and selected the parameters that optimize the sum of Activity × Contact across all other putative elements within 5 Mb of the target gene.

We tested the association of the aggregate MAF <1% variants within each of the aforementioned groups with each lipid trait using continuous traits using the mixed-model SKAT implementation in EPACTS to account for bidirectional effects. We first created group files ("make-group") using annotations from the aforementioned strategies, created VCF-specific kinship matrices ("make-km") using default parameters, and performed association analyses ("group—the mmskat –max-maf 0.01") (http://genome.sph.umich.edu/wiki/EPACTS). Analyses were adjusted for age, sex, cohort, self-reported ethnicity (for MESA), and empircic kinship within each of the VCFs. $P$ values for each analysis were meta-analyzed across the three callsets using Fisher’s method. Statistical significance for each gene-based test was 0.05/20,000 tests = 2.5 x 10^-5.

Lipid extremes analysis. We first defined LDL-C extremes as the top and bottom ancestry-specific 5th percentiles from the data (LDL-C >183 mg/dl or >198.6 mg/dl for EA and AA, respectively; LDL-C <72.9 mg/dl or <71 mg/dl for EA and AA, respectively). We next cataloged mutations in Mendelian genes previously linked to extreme LDL-C (Supplementary Table 13). We included variants that were previously linked to Mendelian dyslipidemia in ClinVar ("pathogenic" or "likely pathogenic" with no "benign") or loss-of-function, and had an allele frequency <1% (autosomal dominant) or <10% (autosomal recessive). Genotypes were only considered based on expected inheritance pattern (autosomal dominant or autosomal recessive). We evaluated three distinct approaches to generate weighted polygenic scores using prior genome-wide association analysis summary statistics: (1) only lead variants at genome-wide significant loci, (2) varying $P$ and $LD^2$ thresholds (defined by 1000G CEU using PLINK), and (3) all variants but adjusting weights according to $P$ and $LD^2$ (by 1000G CEU) with LDpred varying rho. To minimize errors from strand flips, SNPs were excluded. The scores were calculated as additive sums of risk allele counts for included SNPs multiplied by weights (discovery effect estimates for (1) and (2), or adjusted by LDpred for (3)).

LDpred is a Bayesian approach, calculates a posterior mean effect size for each variant based on a prior (association with LDL-C in a previously published study) and subsequent shrinkage based on the extent to which this variant is correlated with similarly associated variants in a reference population. The underlying Gaussian distribution additionally considers the fraction of causal (e.g., non-zero effect sizes) markers. Because this fraction is unknown for any given disease, LDpred uses a chromosome-wide plausible values to construct different polygenic scores.

Polygenic scores were generated within the HUNT cohort, the training set is Lipid values were extracted from the electronic health record; absence of lipid-lowering therapy was prioritized. For each trait, the model with the best fit, as measured by $R^2$, was chosen to apply to the testing set, TOPMed samples.

In a multivariable model, we associated likelihood of membership within the extreme tail of a trait with monogenic mutation carrier status, high (top 5th percentile) or low (bottom 5th percentile) polygenic score, age, age², and sex, separately in European American (EA from FHS and MESA-AA) and African American (AA from JHS and MESA-AA) samples. We also ran linear regression models with continuous LDL-C and the independent variables listed above.

Data availability. Individual whole-genome sequence data for TOPMed whole genomes (FHS, JHS, OOA, and MESA) are available through restricted access via the TOPMed dbGaP Exchange Area. The accession numbers are: FHS phs000974.v1.p1, JHS phs000964.v1.p1, OOA phs000956.v1.p1, and MESA phs001416.v1.p1. Individual-level harmonized lipids used for analysis are also available through restricted access via the TOPMed dbGaP Exchange Area. Summary-level genotype data are available through the BRAVO browser (https://bravo.sph.umich.edu/). The Finnish WGS and array genotype data can be accessed through the THL Biobank (https://thl.fifi.fi/web/thl-biobank). The WGS data at Estonian Genome Center, University of Tartu, can be accessed via the Estonian Biobank (www.biobank.ee).

Received: 15 May 2018 Accepted: 22 June 2018 Published online: 23 August 2018

References

the manuscript. P.N., G.M.P., S.K., and NHLBI TOPMed Lipids Working Group provided administrative, technical or material support.

**Additional information**

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05747-8.

**Competing interests:** The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### NHLBI TOPMed Lipids Working Group


25New York Genome Center, New York, NY 10013, USA. 26Massachusetts General Hospital, Boston, MA 02114, USA. 27Wake Forest Baptist Health, Winston-Salem, NC 27157, USA. 28Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA. 29University of Pennsylvania, Philadelphia, PA 19104, USA. 30Emory University, Atlanta, GA 30322, USA. 31University of Maryland School of Medicine, Baltimore, MD 21201, USA. 32University of Washington, Seattle, WA 98195, USA. 33University of Mississippi, Jackson, MS 38677, USA. 34National Institutes of Health, Bethesda, MD 20892, USA. 35Johns Hopkins University, Baltimore, MD 21218, USA. 36University of Kentucky, Lexington, KY 40506, USA. 37Duke University, Durham, NC 27708, USA. 38University of Alabama, Birmingham, AL 35487, USA. 39Stanford University, Stanford, CA 94305, USA. 40University of Wisconsin Milwaukee, Milwaukee, WI 53211, USA. 41Cleveland Clinic, Cleveland, OH 44195, USA. 42University of Colorado at Denver, Denver, CO 80204, USA. 43Columbia University, New York, NY 10027, USA. 44Boston University, Boston, MA 02215, USA. 45Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife 52011-000, Brazil. 46University of Michigan, Ann Arbor, MI 48109, USA. 47University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78520, USA. 48University of Texas Health, Houston, TX 77225, USA. 49National Jewish Health, Denver, CO 80206, USA. 50Medical College of Wisconsin, Milwaukee, WI 53226, USA. 51University of California, San Francisco, San Francisco, CA 94143, USA. 52Women’s Health Initiative, Seattle, WA 98109, USA. 53University of California, Los Angeles, CA 90095, USA. 54Brigham & Women’s Hospital, Boston, MA 02115, USA. 55University of Virginia, Charlottesville, VA 22903, USA. 56Los Angeles Biomedical Research Institute, Los Angeles, CA 90052, USA. 57The Broad Institute, Cambridge, MA 02142, USA. 58National Taiwan University, 10617 Taipei, Taiwan. 59University of Vermont, Burlington, VT 05405, USA. 60Bloom Systems Research Institute UCEFS, San Francisco, CA 94118, USA. 61University of Illinois at Chicago, Chicago, IL 60607, USA. 62Mayo Clinic, Rochester, MN 55905, USA. 63Vanderbilt University, Nashville, TN 37235, USA. 64University of Cincinnati, Cincinnati, OH 45220, USA. 65Icahn School of Medicine at Mount Sinai, New York 10029 NY, USA. 66University of North Carolina, Chapel Hill, NC 27599, USA. 67University of Texas Rio Grande Valley School of Medicine, Edinburg, TX 78539, USA. 68Washington University in St. Louis, St. Louis, MO 63103, USA. 69Brown University, Providence, RI 02912, USA. 70Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. 71University of Pittsburgh, Pittsburgh, PA 15260, USA. 72Yale University, New Haven, CT 06520, USA. 73University of Texas Rio Grande Valley School of Medicine, San Antonio, TX 78229, USA. 74Tulane University, New Orleans, LA 70118, USA. 75University of Iowa, Iowa City, IA 52242, USA. 76National Health Research Institute Taiwan, 350 Zhunan Township, Taiwan. 77Blood Works Northwest, Seattle, WA 98106, USA. 78Taichung Veterans General Hospital, Taichung, Taiwan. 79Ohio State University Wexner Medical Center, Columbus, OH 43210, USA. 80Bloom Works Northwest, Seattle, WA 98106, USA. 81NIH National Heart, Lung, and Blood Institute, Bethesda, MD 20892, USA. 82Albert Einstein College of Medicine, New York, NY 10461, USA. 83Blood Works Northwest, Seattle, WA 98104, USA. 84 Loyola University, Maywood, IL 60153, USA. 85Harvard School of Public Health, Boston, MA 02115, USA. 86George Washington University, Washington, DC 20052, USA. 87Harvard University, Cambridge, MA 02138, USA. 88University of Arizona, Tucson, AZ 85721, USA. 89Howard University, Washington, DC 20059, USA. 90University at Buffalo, Buffalo, NY 14260, USA. 91University of Minnesota, Minneapolis, MN 55455, USA. 92Northwestern University, Chicago, IL 60208, USA. 93Blood Works Northwest, Seattle, WA 98107, USA. 94Harvard Medical School, Boston, MA 02115, USA. 95University of Colorado, Denver, CO 80204, USA. 96University of California, Los Angeles, CA 90095, USA. 97University of Colorado at Boulder, Boulder, CO 80309, USA. 98Henry Ford Health System, Detroit, MI 48202, USA.